

Nonparametric functionals as generalized functions

Victoria Zinde-Walsh *

McGill University and CIREQ

victoria.zinde-walsh@mcgill.ca

(514) 398 4834

March 7, 2013

*The support of the *Fonds québécois de la recherche sur la société et la culture* (FRQSC) is gratefully acknowledged.

Running head: Nonparametric functionals

Victoria Zinde-Walsh

Department of Economics, McGill University

855 Sherbrooke Street West,

Montreal, Quebec, Canada

H3A 2T7

Abstract

The paper considers probability distribution, density, conditional distribution and density and conditional moments as well as their kernel estimators in spaces of generalized functions. This approach does not require restrictions on classes of distributions common in nonparametric estimation. Density in usual function spaces is not well-posed; this paper establishes existence and well-posedness of the generalized density function. It also demonstrates root-n convergence of the kernel density estimator in the space of generalized functions. It is shown that the usual kernel estimator of the conditional distribution converges at a parametric rate as a random process in the space of generalized functions to a limit Gaussian process regardless of pointwise existence of the conditional distribution. Conditional moments such as conditional mean are also be characterized via generalized functions. Convergence of the kernel estimators to the limit Gaussian process is shown to hold as long as the appropriate moments exist.

1 Introduction

A probability distribution function, F , that corresponds to a Borel measure on a Euclidean space R^k (or its subspace) is always defined in the space of bounded functions. It can be viewed as the right-hand side of an integral equation:

$$I(f) = F; \tag{1}$$

where the density represents the solution to the inverse problem

$$f = \partial^k F. \tag{2}$$

Here I represents an integration operator for R^k : $I(f)(x) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_k} f(w) dw_1 \dots dw_k$ and $\partial^k = \frac{\partial^k}{\partial x_1 \dots \partial x_k}$ the inverse differentiation operator.

When does the solution to the inverse problem exist?

In the usual approach the integral operator I is assumed to operate on the space of integrable functions, e.g. L_1 (absolutely integrable functions) or L_2 (square integrable functions), - see e.g. Devroye and Györfi (1985), Carrasco, Florens, Renault (2007). The operator I maps density functions in L_1 into the space of absolutely continuous distribution functions. In this case the inverse operator ∂^k is defined and the inverse problem has a unique solution.

The property of well-posedness requires that the solution continuously depend on the right-hand side function, in other words, if distribution func-

tions are close, the corresponding densities should be close as well. However, in spaces of integrable functions the inverse problem is not well-posed: while the operator I is continuous on L_1 (or another L_p space) the inverse operator ∂^k is not. The example below (from Zinde-Walsh, 2011) illustrates lack of well-posedness.

Example. Consider the space $D([0, 1])$ of univariate absolutely continuous distribution functions on the interval $[0, 1]$ in the uniform metric: the distance between two distributions, F_1, F_2 is $d(F_1, F_2) = \max_{x \in [0, 1]} |F_1(x) - F_2(x)|$; this is the image space of the operator $I(\cdot)$ defined on $L_1([0, 1])$.

Denote by $[v]$ the integer part of v , that is the largest integer that is $\leq v$. Let $I(x \in A)$ denote the indicator function of set A , that equals 1 if x is in A , zero otherwise. With $\varepsilon = \frac{\bar{\varepsilon}}{2}$ define densities

$$\begin{aligned} f_1(x) &= 2 \sum_{m=0}^{\left[\frac{\varepsilon^{-1}+1}{2}\right]-1} I(x \in [2m\varepsilon, (2m+1)\varepsilon)); \\ f_2(x) &= 2 \sum_{m=0}^{\left[\frac{\varepsilon^{-1}+1}{2}\right]-1} I(x \in [(2m+1)\varepsilon, (2m+2)\varepsilon)). \end{aligned}$$

The densities f_1 and f_2 have supports that do not intersect, it is easily seen that at each point they differ by 2: $|f_1(x) - f_2(x)| = 2$; it follows that the $L_1([0, 1])$ difference between them is 2. The corresponding distributions are $F_1 = I(f_1)$ and $F_2 = I(f_2)$. It is easy to establish by integration that

$$\max_{x \in [0, 1]} |F_1(x) - F_2(x)| \leq 2\varepsilon = \bar{\varepsilon},$$

and thus the inverse operator is not continuous.

Thus although a solution to the inverse problem in the L_1 space exists for absolutely continuous distributions, the problem is not well-posed.

By contrast, in the appropriate space of generalized functions the solution to the density problem exists without any restrictions on the distribution function and is well-posed; as proved in section 2 below this follows from the known properties of generalized functions. The fact that generalized functions can be useful when non-differentiability prevents the use of Taylor expansions was discussed e.g. in Phillips (1991) for LAD estimation, and continued in some econometric literature that followed.

The statistical inverse problem is solved often with a kernel density estimator. Consider a random sample of observations from a distribution F , $\{x_i\}_{i=1}^n$, $x_i \in R^k$. With a chosen kernel function, K and bandwidth (vector) h the estimator is

$$\widehat{f(x)} = \frac{1}{n \prod_{j=1}^k h_j} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right), \quad (3)$$

where h has components h_1, \dots, h_k and $K(\frac{x_i - x}{h})$ is a multivariate function with the argument $(\frac{x_i - x}{h}) = \left(\left(\frac{x_{i1} - x_1}{h_1}\right), \dots, \left(\frac{x_{ik} - x_k}{h_k}\right)\right)$. We shall proceed with the following assumption on the kernel.

Assumption 1 (kernel).

- (a). $K(w)$ is an ordinary bounded function on R^k ; $\int K(w)dw = 1$;
- (b). Support of K belongs to $[-1, 1]^k$;

(c). $K(w)$ is an l -th order kernel: for $w = (w_1, \dots, w_k)$ the integral

$$\int w_1^{j_1} \dots w_k^{j_k} K(w) dw_1 \dots dw_k \begin{cases} = 0 & \text{if } j_1 + \dots + j_k < l; \\ < \infty & \text{if } j_1 + \dots + j_k = l. \end{cases}$$

The finite support and boundedness assumptions can be relaxed and are introduced to simplify assumptions and derivations; K is not restricted to be symmetric or non-negative.

Denote by \bar{K} the integral of the kernel function, then

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^n \bar{K}\left(\frac{x_i - x}{h}\right) \quad (4)$$

is an estimator of the distribution function, $F(x)$. The properties of these estimators depend on K and h and are well established (Azzalini, 1981). Generally for $h \rightarrow 0$ as $n \rightarrow \infty$ with $nh \rightarrow \infty$, $\widehat{F}(x)$ is a root- n consistent and asymptotically Gaussian estimator of $F(x)$ at any point of continuity; the uniform norm of the difference, $\sup \left| \widehat{F}(x) - F(x) \right|$, converges to zero.

Known convergence properties of $\widehat{f}(x)$ are more complicated; they rely on assumptions about the existence and smoothness of the density, $f(x)$; the convergence rate is slower than root- n and depends on the order of the kernel and the rate of the bandwidth $h \rightarrow 0$ (Pagan and Ullah, 1999). As shown in Examples 3-5 in Zinde-Walsh (2008), the estimator $\widehat{f}(x)$ fails to converge pointwise if the distribution is not absolutely continuous (e.g. at a mass point or for a fractal measure); of course, in those cases density itself

cannot be defined pointwise and exists only as the solution, f in (2) to the inverse problem in the space of generalized functions.

When considered in the space of generalized functions the estimators, \hat{f} , are viewed as random continuous linear functionals on spaces of well-behaved functions where convergence to generalized derivatives of distribution functions (solutions to the inverse problem) can be established without any assumptions on the underlying distribution. Moreover, convergence of kernel estimators can be faster and even at parametric rates. This result has features common to other results on convergence of random functionals of density as discussed, e.g. in Anderson et al (2012) and is derived here in section 3. This result relies on the derivation of the rate of bias in generalized functions that was provided in Zinde-Walsh (2008) but gives the derivation of the covariance functional that corrects the one in that paper.

Conditioning is somewhat awkward and there are many different ways to streamline the representation of conditional measures and distribution functions (Chiang and Pollard, 1997, Pfanzagl, 1979 among others). Here we focus on the distribution function $F(x, y)$ on $R^{d_x} \times R^{d_y}$ and distribution of $y \in R^{d_y}$ conditional on $x \in R^{d_x}$. In this case typically the conditional distribution $F_{y|x}$ function is represented via a fraction $\frac{\partial^{d_x} F(x, y)}{f_x(x)}$, where the differentiation operator is applied to the x argument of $F(x, y)$ and $f_x(x)$ represents the density of the marginal distribution. Of course such a representation makes stringent requirements on the smoothness of the appropriate functions. Here the case of an arbitrary continuous conditioning distribution

is considered without requiring differentiability; it is shown that for this case the conditional distribution and conditional density have a straightforward representation as generalized functions on appropriate spaces. The representation is in terms of functionals involving the conditioning distribution (rather than the conditioning variable) as an argument; this representation avoids the nonlinearity introduced by the denominator. When the usual representation holds, a simple correspondence between the two representations is established. Conditional density, $f_{y|x}$ is defined as a generalized derivative of the conditional distribution generalized function.

The convergence of the usual kernel estimator of the conditional distribution is known under smoothness assumptions (Pagan and Ullah, 1999, Li and Racine, 2007) and utilizes the properties of the kernel density estimator; the density appears in the denominator of the statistic requiring some support assumptions and possibly regularization to converge. Here the root-n convergence of the kernel estimator to a limit Gaussian process in generalized function space is established without any extra restrictions on the distribution.

An interpretation of a conditional moment function is provided here in the space of generalized functions, thus again without any restriction beyond continuity of conditioning distribution. For estimators, such as for conditional mean kernel estimator the asymptotic properties are established, the result is then that root-n convergence in generalized functions obtains for the kernel estimator without any restrictions on smoothness of distribution

functions.

The theoretical results of this paper extend the usual representation of the density, conditional distribution and density and conditional moments to situations where these may not exist in an ordinary sense. The advantage that this approach provides is its generality. On the other hand, the topology in the spaces of generalized functions is weak and well-posedness does not imply convergence in norm.

The asymptotic results provide a general approach, so that when the usual assumptions may fail there is still a sense in which consistency holds. Moreover a root-n convergence rate obtains, again as a consequence of the weak topology with no guarantee of good convergence in norm. The practical advantage is in the possibility of utilizing the generalized random process and its limit process for inference without making any restrictions on the distribution.

2 Density as solution to a well-posed inverse problem in the space of generalized functions

For the definitions and results pertaining to spaces of generalized functions the main references are to books by Schwartz (1966) Gel'fand and Shilov (1964). A useful summary is in Zinde-Walsh (2008, 2012); the main defini-

tions follow.

Consider a space of well-behaved "test" functions, $D_\infty(R^k)$ of infinitely differentiable functions with bounded support, or any of the spaces $D_m(R^k)$ of m times continuously differentiable functions (with bounded support); sometimes the domain of definition can be an open subset W of R^k , typically here $W = (0, 1)^k$. Denote the generic space by $D(W)$; convergence in $D(W)$ is defined as follows: a sequence $\psi_n \in D(W)$ converges to zero if all ψ_n are defined on a common bounded support in W and ψ_n as well as all the l -th order derivatives (with $l \leq m$ for D_m or all $l < \infty$ for D_∞) converge pointwise to zero. The space of generalized functions is the dual space, D^* , the space of linear continuous functionals on $D(W)$ with the weak topology: a sequence of elements of D^* converges if the sequence of values of the functionals converges for any test function from $D(W)$. The usual notation is to write the value of the functional f applied to a test function $\psi \in D(W)$ as (f, ψ) ; then a sequence f_n converges to f if for any ψ convergence $(f_n, \psi) \rightarrow (f, \psi)$ holds.

Assume that functions in $D(W)$; $W \subseteq R^k$ are suitably differentiable, e.g. at least k times continuously differentiable. Then for any $\psi \in D(W)$, and $F \in D^*$ define a generalized derivative $f \in D^*$; $f = \frac{\partial^k}{\partial x_1 \dots \partial x_k} F$ as the functional with values given by:

$$(f, \psi) = (-1)^k (F, \frac{\partial^k \psi}{\partial x_1 \dots \partial x_k}). \quad (5)$$

If the right-hand side is expressed via a regular locally summable function

as is the case when F is a probability distribution function, then it can be computed by integration:

$$\left(F, \frac{\partial^k \psi}{\partial x_1 \dots \partial x_k}\right) = \int \dots \int F(x_1, \dots, x_k) \frac{\partial^k \psi(x_1, \dots, x_k)}{\partial x_1 \dots \partial x_k} dx_1 \dots dx_k.$$

For the function F (5) the functional on the right-hand side defines the generalized derivative: $f = \frac{\partial^k F}{\partial x_1 \dots \partial x_k}$.

First consider density as a generalized function on the space $D_\infty(W)$.

Theorem 1. *The inverse problem (1) for any cumulative probability distribution function F has the solution f defined by (5) in the space of generalized functions D^* for $D_\infty(W)$. The problem is well-posed. When density exists as an integrable function, $f(x)$, it provides the generalized function f via the value of the corresponding functional:*

$$(f, \psi) = \int \dots \int f(x_1, \dots, x_k) \psi(x_1, \dots, x_k) dx_1 \dots dx_k. \quad (6)$$

Proof.

Any distribution function F on R^k is a monotone bounded function and as such is locally integrable on any bounded set; a function like that represents a regular element in the space of generalized functions, D^* , for $D_\infty(W)$ defined above. Then (5) defines f as the generalized derivative of F , the generalized density function.

The differentiation operator $\partial^k = \frac{\partial^k}{\partial x_1 \dots \partial x_k}$ on the space of generalized functions D^* is defined for any regular function and is a continuous operator

(Schwartz, p.80). Thus the solution f continuously depends on F in these spaces providing well-posedness.

If density f exists as a regular integrable function, its integral coincides with the function F and integration by parts of (6) provides (5). Thus f , the solution to the inverse problem in the space D^* is consistent with the solution when it exists as an ordinary function. ■

Corollary. *The result of the Theorem applies in the space of generalized functions on $D_m(W)$, $m \geq k$.*

Proof.

Indeed, consider the space $D_\infty(W) \subset D_k(W)$. By the theorem the inverse problem provides the density function f defined as a linear continuous functional on $D_\infty(W)$ via (5). We can extend the functional f to $D_k(W)$ as a linear continuous functional. First note that since F is a regular locally integrable function it represents an element in D_k^* ; then define the functional in D_k^* by (5) for any $\psi \in G_k$, denote it \tilde{f} to distinguish from f defined on $D_\infty(W)$. This \tilde{f} represents a linear continuous functional, so an element in D_k^* . There is an injective mapping of linear topological spaces $D_k^* \rightarrow D_\infty^*$ (Sobolev, 1992 ; in notation there $C^{(k)\#} \rightarrow C^{(\infty)\#}$), thus by this mapping \tilde{f} maps into f and the inverse problem is solved in D_k^* and is well-posed there. ■

3 Gaussian limit process for the kernel density estimator in the space of generalized functions

We now describe the limit process for the kernel estimator (3) as $\bar{h} = \max_{1 \leq j \leq k} h_j \rightarrow 0$ with $n \rightarrow \infty$, as a generalized random process. Such a description was in Zinde-Walsh (2008), but there was an error in the variance computation that is corrected here. The main result here is that in the generalized functions space convergence of the kernel density estimator can be at a parametric rate for a suitable selection of the kernel and bandwidth; unlike the usual case in the literature this selection alone provides the result independently of any properties (smoothness) of the distribution.

Recall that convergence of generalized random functions is defined (see, e.g. Gel'fand and Vilenkin, 1964 or summary in Zinde-Walsh, 2008) as weak convergence of random linear continuous functionals on the space D . (for any of the D_k, D_∞ , etc. spaces here) that are indexed by the functions in D : stochastic convergence of random functionals, \hat{f} , follows from stochastic convergence of random vectors of values of the functional $\left(\left(\hat{f}, \psi_1 \right), \dots, \left(\hat{f}, \psi_m \right) \right)'$ for any finite set (ψ_1, \dots, ψ_m) with $\psi_l \in D$. Thus we need to consider the behavior of such random vectors.

Theorem 2 in Zinde-Walsh (2008) gives the convergence rate $O(\bar{h}^l)$ for the generalized bias function of the kernel estimator based on a random sample

and the expression for the bias for $\psi \in D_{l+k}$ and kernel K of order l :

$$E\hat{f} - f \approx O(\bar{h}^l),$$

more specifically for any ψ the bias functional provides $(E\hat{f}, \psi) - (f, \psi) =$

$$(-1)^l \sum_{\Sigma m_i = l} \int \prod_{i=1}^k \frac{h_i^{m_i}}{m_i!} F(\tilde{x}) \frac{\partial^{l+k} \psi}{\partial x_1^{m_{1i}+1} \dots \partial x_k^{m_{ki}+1}}(\tilde{x}) d\tilde{x} \int K(w) w_1^{m_1} \dots w_k^{m_{ki}} dw + R(h),$$

where $R(h) = o(\bar{h}^l)$; if $\psi \in D_{l+k+1}$ then $R(h) = O(\bar{h}^{l+1})$. Note that $(f, \psi) = E(\psi)$ where expectation is with respect to the measure given by F .

Denote the expression

$$(-1)^l \sum_{\Sigma m_i = l} \int \prod_{i=1}^k \frac{(h_i/\bar{h})^{m_i}}{m_i!} F(\tilde{x}) \frac{\partial^{l+k} \psi}{\partial x_1^{m_{1i}+1} \dots \partial x_k^{m_{ki}+1}}(\tilde{x}) d\tilde{x} \int K(w) w_1^{m_1} \dots w_k^{m_{ki}} dw$$

by $(B(h, K), \psi)$ as it represents the value of a linear continuous functional $B(h, K)$ applied to ψ . The $B(h, K)$ is the leading term in the generalized bias function for the kernel estimator:

$$Bias(\hat{f}) = E\hat{f} - f = \bar{h}^l B(h, K) + o(\bar{h}^l); \quad (8)$$

where for any $\psi \in D_{l+k+1}$

$$(E\hat{f}, \psi) - (f, \psi) = \bar{h}^l (B(h, K), \psi) + o(\bar{h}^l).$$

The following Theorem gives the limit process for the kernel estimator of density.

Theorem 2. *For a kernel function K satisfying Assumption A, if $\bar{h} \rightarrow 0$ and $\bar{h}^{2l}n = O(1)$ as $n \rightarrow \infty$ the sequence of generalized random processes $n^{\frac{1}{2}} \left(\hat{f} - f - \bar{h}^l B(h, K) \right)$ converges to a generalized Gaussian process with mean functional zero and covariance functional C which for any $\psi_1, \psi_2 \in D_{l+k}$ provides*

$$(C, (\psi_1, \psi_2)) = E([\psi_1(x) - E\psi_1(x)][\psi_2(x) - E\psi_2(x)]) = \text{cov}(\psi_1, \psi_2). \quad (9)$$

If $n\bar{h}^{2l} \rightarrow 0$, then $\hat{f} - f$ converges at the parametric rate \sqrt{n} to a generalized zero mean Gaussian process with covariance functional C in (9).

Proof. See appendix.

The condition on the bandwidth that makes it possible to eliminate the bias asymptotically is less stringent than in the usual topologies and also than that originally stated in Zinde-Walsh (2008). Under this requirement on the bandwidth convergence is actually at a parametric rate and the limit covariance does not involve the kernel function.

4 Distribution function conditional on some

variables and conditional density in the space of generalized functions

Conditioning is an awkward operation as discussed e.g. in Chang and Pollard (1997). Here the question posed is limited to conditioning on a variable or vector in a joint distribution, that is given a joint distribution function $F_{x,y}(\cdot, \cdot)$ on $R^{d_x} \times R^{d_y}$ define a (generalized) function $F_{y|x}(\cdot, \cdot)$ that represents the conditional distribution of y given x . A problem associated with such conditioning is that the conditional distribution function may not exist for every point x .

Denote by F_x, F_y the marginal distribution functions of x, y , correspondingly.

Consider limits of ratios to define conditioning:

$$F_{y|x} = \lim_{\Delta \rightarrow 0} \frac{F_{x,y}(x + \Delta, y) - F_{x,y}(x, y)}{F_x(x + \Delta) - F_x(x)}. \quad (10)$$

As discussed in numerous papers there is a problem defining such a limit (e.g. Pfazagle, 1979); here it will be demonstrated that the limit exists in a particular space of generalized functions. Assume that the distribution function F_x is continuous; continuity of this distribution of course does not preclude singularity.

Assumption 2. *The marginal distribution function $F_x(x)$ is continuous on R^{d_x} .*

Note that although support of the random y belongs to R^{d_y} it could be a discrete set of points, thus we do not restrict y to be continuously distributed.

Consider the copula function (Sklar, 1973): $C_{F_x, F_y}(a, b)$ on $W = (0, 1)^2$ that is identical to the joint distribution function, that is for the mapping $M : R^{d_x} \times R^{d_y} \rightarrow W$ defined by $\{x, y\} \rightarrow \{F_x(x), F_y(y)\}$ we get the corresponding mapping $M^*(F_{x,y}(x, y)) = C_{M(x,y)}(M(x, y))$ with

$$C_{M(x,y)}(M(x, y)) = C_{F_x, F_y}(F_x(x), F_y(y)) = F_{x,y}(x, y).$$

Thus (10) is equivalent to

$$F_{y|x} = \lim_{\Delta \rightarrow 0} \frac{C_{F_x, F_y}(F_x(x + \Delta), F_y(y)) - C_{F_x, F_y}(F_x(x), F_y(y))}{F_x(x + \Delta) - F_x(x)},$$

denote $F_x(x + \Delta) - F_x(x)$ by $\tilde{\Delta}$, then by Assumption 2, continuity of F_x , $\Delta \rightarrow 0$ implies $\tilde{\Delta} \rightarrow 0$ thus the limit is equivalent to

$$\lim_{\tilde{\Delta} \rightarrow 0} \frac{C_{F_x, F_y}(a + \tilde{\Delta}, b) - C_{F_x, F_y}(a, b)}{\tilde{\Delta}}.$$

Since with respect to its second argument the copula function and the limit are ordinary functions we concentrate on being able to define the generalized derivative with respect to the first argument. In particular, for any $\psi \in D(W)$, given the second argument the value of the functional $\left((C_{F_x, F_y})'_1, \psi\right) = -(C_{F_x, F_y}, \psi')$. This implies that we can define the value

of the functional $F_{y|x}$ on $D(W)$ by

$$(F_{y|x}, \psi) = -(C_{F_x, F_y}, \psi') = - \int F_{x,y}(x, y) \psi'(F_x(x)) dF_x(x). \quad (11)$$

Thus we can define the conditional distribution $F_{y|x}$ as a generalized function in the space $D^*(W)$.

When $d_x = 1$ this is an exhaustive representation. When $d_x > 1$ it may be advantageous to consider a derivative with respect to a d_x -dimensional argument. Consider the conditioning vector, x , component-wise, and consider the multivariate copula function, $C_{F_{x_1}, \dots, F_{x_d}, F_y}(F_{x_1}, \dots, F_{x_d}, F_y)$; to simplify notation we drop the subscript to denote it simply by C . Then by a similar argument for any $\psi \in D(W)$ where $W = (0, 1)^{d_x}$ we obtain

$$(F_{y|x}, \psi) = (-1)^{d_x} (C, \partial^{d_x} \psi) =$$

$$(-1)^{d_x} \int \dots \int F_{x,y}(x, y) \partial^{d_x} \psi(F_{x_1}(x_1), \dots, F_{x_{d_x}}(x_{d_x})) dF_{x_1}(x_1) \dots dF_{x_{d_x}}(x_{d_x}). \quad (12)$$

Remark 1. Similarly to Corollary 1, the generalized function $F_{y|x}$ can be extended as a linear continuous functional from being defined on the space $D(W)$ of infinitely differentiable functions to a linear continuous functional defined by (11) on any space $D_k(W)$ with $k \geq 1$ and for (??) to $D_k(W)$ for the corresponding W and $k \geq d_x$.

Remark 2. If the function C were suitably differentiable the functional $(F_{y|x}, \psi)$ would be defined for any continuous ψ with bounded support, that

is on the space $D_0(W)$ by $(\partial^{d_x} C(\dots, \cdot), \psi) :$

$$(F_{y|x}, \psi) = \int \dots \int \partial^{d_x} C(F_{x_1}, \dots, F_{x_{d_x}}, F_y) \psi(F_{x_1}, \dots, F_{x_{d_x}}) dF_{x_1} \dots dF_{x_{d_x}}. \quad (13)$$

In the y argument the conditional distribution is an ordinary function so here y is considered just as a parameter of the generalized function. However, the definition of $F_{y|x}$ in (11) can be extended to a functional for functions defined on the product space; for any $\psi_{x,y} = \psi_x(x_1, \dots, x_{d_x}) \psi_y(y_1, \dots, y_{d_y}) \in D((0, 1)^{d_x}) \times D(R^{d_y})$ define the value of the functional by $(F_{y|x}, \psi_{x,y}) =$

$$(-1)^{d_x} \int \dots \int F(x, y) \partial^{d_x} \psi_x(F_{x_1}, \dots, F_{x_{d_x}}) \psi_y(y_1, \dots, y_{d_y}) dF_{x_1} \dots dF_{x_{d_x}} dy_1 \dots dy_{d_y}.$$

To define conditional density $f_{y|x}$ as a generalized function one would have

$$(f_{y|x}, \psi_{x,y}) = (-1)^{d_x+d_y} \int \dots \int F_{x,y}(x, y) \partial^{d_x} \psi_x(F_{x_1}, \dots, F_{x_{d_x}}) \partial^{d_y} \psi_y(y_1, \dots, y_{d_y}) dF_{x_1}(x_1) \dots dF_{x_{d_x}}(x_{d_x}) dy_1 \dots dy_{d_y}. \quad (14)$$

In general, the conditional distribution and conditional density depend on the conditioning variables, x , via the marginals, F_x ; considering generalized functions makes this explicit.

There are cases when the conditional distribution and conditional density are defined on the Euclidean space R^{d_x} . This is possible if the distribution function F_x is strictly monotone in each argument; then the corresponding generalized density function is positive, moreover, since a monotone function

is a.e. differentiable, $\partial^{d_x} F_{x,y}(x, y)$ and $f_x(x) = \partial^{d_x} F_x(x)$ exist a.e. and $f_x(x) > 0$. When the density f_x is a continuous function the conditional distribution can be represented as a functional on a function space on R^{d_x} that can be derived from the general representation above in $D^*(W)$.

Indeed, any distribution function, $F(x, y)$, where we focus on the argument x , via the copula representation can be considered as a functional on $D(W)$. Let Φ denote the class of such distribution functions, then $\Phi \subset D^*(W)$. Moreover the representation (12) demonstrated that any conditional distribution $F_{|x}(x, y)$ also defines a linear continuous functional on $D(W)$. Denoting by $\Phi_{|x}$ the class of conditional distributions we thus have shown that $\Phi_{|x} \subset D^*(W)$. By the remark, we can relax the differentiability conditions and consider $\Phi_{|x} \subset D_k^*(W)$; when the distribution function is differentiable in x , we set $k = 0$. On the other hand, then a continuous density function, $f_x > 0$ exists and the conditional distribution can be represented by an ordinary function $\frac{\partial^{d_x} F_{x,y}(x,y)}{f_x(x)}$; denote by Φ_c the class of distributions that are continuously differentiable in x with $f_x > 0$ on R^{d_x} , and by $\Phi_{c|x}$ the class of corresponding conditional distributions. Then $\Phi_c \subset D_0^*(R^{d_x})$ and as well $\Phi_{c|x} \subset D_0^*(R^{d_x})$, where the space $D_0^*(R^{d_x})$ is the space of continuous functions with bounded support in R^{d_x} . Since $\Phi_{c|x} \subset \Phi_{|x}$, any conditional distribution that exists in the ordinary sense and thus is in $\Phi_{c|x}$, has two representations: one as a functional on $D_0(W)$ defined above and the second

as a functional on $D_0(R^{d_x})$ that provides for any $\tilde{\psi} \in D_0(R^{d_x})$

$$(F_{y|x}, \tilde{\psi}) = \int \dots \int \frac{\partial^{d_x} F_{x,y}(x, y)}{f_x(x)} \tilde{\psi}(x) dx_1 \dots dx_{d_x}. \quad (15)$$

The following lemma shows that the two representations are compatible and each can be easily obtained from the other.

Lemma. *Suppose that $F_{x,y} \in \Phi_c$. Then the value of the functional given by (13) for $\psi \in D_0(0, 1)^{d_x}$ is the same as the value of the functional given by (15) for $\tilde{\psi}(x) = f_x(x)\psi(F(x)) \in D_0(R^{d_x})$; and vice versa: given (15) the value of (13) for $\psi(F_{x_1}, \dots, F_{x_{d_x}}) = \frac{\tilde{\psi}(x_1, \dots, x_{d_x})}{f_x(x_1, \dots, x_{d_x})}$, where x_i is uniquely determined by the value of F_{x_i} : $x_i = F_{x_i}^{-1}(F_{x_i}(x_i))$, is the same.*

Proof. For any $\psi \in D(0, 1)^{d_x}$ define $\tilde{\psi}$ on R^{d_x} by $\tilde{\psi}(x) = f_x(x)\psi(F(x))$, then $(F_{y|x}, \psi)$ defined by (12) by differentiability of $F_{x,y}$ in x is equal to

$$(F_{y|x}, \tilde{\psi}) = \int \dots \int \frac{\partial^{d_x} F_{x,y}(x, y)}{f_x(x)} \tilde{\psi}(x) dx_1 \dots dx_{d_x}.$$

Denote by z_i the value $F_{x_i}(x)$, $i = 1, \dots, d_x$; then (for clarity we subscript the operator ∂ by the variable(s) with respect to which we differentiate):

$$\partial_z^{d_x} F_{x,y} \left(F_{x_1}^{-1}(z_1), \dots, F_{x_{d_x}}^{-1}(z_{d_x}), y \right) f_x(x) = \partial_x^{d_x} F_{x,y}(x, y).$$

The r.h.s. of (12) provides

$$\begin{aligned}
& (-1)^{d_x} \int \dots \int F_{x,y} \left(F_{x_1}^{-1}(z_1), \dots, F_{x_{d_x}}^{-1}(z_{d_x}), y \right) \partial_z^{d_x} \psi(z_1, \dots, z_{d_x}) dz_1 \dots dz_{d_x} \\
&= \int \dots \int \partial_z^{d_x} F_{x,y} \left(F_{x_1}^{-1}(z_1), \dots, F_{x_{d_x}}^{-1}(z_{d_x}), y \right) \psi(z_1, \dots, z_{d_x}) dz_1 \dots dz_{d_x} \\
&= \int \dots \int \frac{\partial_x^{d_x} F_{x,y}(x, y)}{f_x(x)} \psi(F_{x_1}(x_1), \dots, F_{x_{d_x}}(x_{d_x})) f_x(x) dx_1 \dots dx_{d_x}, \\
&\text{and writing this in more concise notation} \\
&= \int \frac{\partial_x^{d_x} F_{x,y}(x, y)}{f_x(x)} \psi(F(x)) f_x(x) dx = \int \frac{\partial_x^{d_x} F_{x,y}(x, y)}{f_x(x)} \tilde{\psi}(x) dx.
\end{aligned}$$

Since f_x is continuous, then $\tilde{\psi}(x) = \psi(F(x)) f_x(x)$ is continuous on R^{d_x} .

For an arbitrary $\tilde{\psi} \in D_0(R^{d_x})$ consider

$$\left(F_{y|x}, \tilde{\psi} \right) = \int \frac{\partial_x F_{x,y}(x, y)}{f_x(x)} \tilde{\psi}(x) dx_1 \dots dx_{d_x}.$$

Do the transformation, then

$$\left(F_{y|x}, \tilde{\psi} \right) = \int \partial_z F_{x,y}(F_x^{-1}(z), y) \frac{\tilde{\psi}(F_x^{-1}(z))}{f_x(F_x^{-1}(z))} dz.$$

Define a continuous function $\psi(F_{x_1}, \dots, F_{x_{d_x}}) = \frac{\tilde{\psi}(x_1, \dots, x_{d_x})}{f_x(x_1, \dots, x_{d_x})}$ on $(0, 1)^{d_x}$, then this equals (13).

■

Suppose now that F_x is absolutely continuous with continuous density function, $f_{y|x}$; then the support of the density function is an open set in R^{d_x} , $S_{y|x}$. The Lemma applies by considering $\tilde{\psi}(x) = f_x(x) \psi(F(x)) \in D_0(S_{y|x})$

in place of $D_0(R^{d_x})$.

5 Limit properties of kernel estimators of conditional distribution in generalized functions

Consider the usual kernel estimator of conditional distribution; typically its limit properties are available under smoothness conditions on the distribution (see, e.g. Li and Racine, 2007). Here the estimator is examined in the space of generalized functions without any restrictions placed on the distribution beyond Assumption 2 (continuity of F_x).

Recall the usual kernel estimator of conditional distribution:

$$\hat{F}_{y|x}(x, y) = \frac{\Sigma \bar{G}\left(\frac{y-y_i}{h_y}\right) K\left(\frac{x_i-x}{h}\right)}{\Sigma K\left(\frac{x_i-x}{h}\right)} \quad (16)$$

$$= \frac{\frac{1}{n} \Sigma \bar{G}\left(\frac{y-y_i}{h_y}\right) \frac{1}{h^{d_x}} K\left(\frac{x_i-x}{h}\right)}{\hat{f}_x(x)}, \quad (17)$$

where \bar{G} is the integral of a kernel function G similar to K that satisfies Assumption 1 on R^{d_y} and K satisfies Assumption 1 on R^{d_x} . Sometimes \bar{G} is assumed to be the indicator function $I(w > 0)$.

To simplify exposition we assume that each component of vector x is associated with the same (scalar) bandwidth parameter h ; it is not difficult to generalize to the case of distinct bandwidths.

Theorem 3. *Suppose that Assumption 1 on the kernel K and either*

a similar assumption for G holds, or \bar{G} is the indicator function, the bandwidth parameter $h = cn^{-\alpha}$, where $\alpha < \frac{1}{4}$ and Assumption 2 holds. Then for a random sample $\{(x_i, y_i)\}_{i=1}^n$ the estimator $\hat{F}_{y|x}(x, y)$ as a generalized random function on $D(W)$ converges to the conditional distribution generalized function $F_{y|x}$ defined by (11) at the rate $n^{-\frac{1}{2}}$; the limit process for $\sqrt{n}(\hat{F}_{y|x} - F_{y|x})$ on $D(W)$ is given by a $\psi \in D(W)$ indexed random functional, $Q_{y|x}$ with $(Q_{y|x}, \psi) =$

$$(-1)^{d_x} \left[\int F_{xy}(\partial^{d_x} \partial^{d_x} \psi)(F_x) U_x dF_x + \int F_{xy}(\partial^{d_x} \psi)(F_x) dU_x + \int (\partial^{d_x} \psi)(F_x) U_{xy} dF_x \right],$$

where U_x, U_{xy} are Brownian bridge processes with dimension $d_x, d_y + d_x$, correspondingly; as a generalized random process the limit process $Q_{y|x}$ of $\sqrt{n}(\hat{F}_{y|x} - F_{y|x})$ is Gaussian with mean functional zero and covariance bilinear functional C , given for any ψ_1, ψ_2 by

$$(C, (\psi_1, \psi_2)) = \text{cov}[(Q_{y|x}, \psi_1), (Q_{y|x}, \psi_2)].$$

Proof. See Appendix.

This result is general in that the root-n convergence holds here regardless of whether the marginal density exists. If it does exist the result could be restated for conditional distribution as a generalized function on $D_0(R^{d_x})$ by (15).

Remark 3. Sometimes for a singular distribution the kernel estimator $\hat{f}_x(x)$ diverges at a specific rate, as e.g. in Lu (1999) where at points x

in support of density $\hat{f}_x(x) = h^{d-1}b + o_p(h^{d-1})$ with some $b > 0$ and $d = \frac{\ln 2}{\ln 3} < 1$. In the univariate case this is discussed in Example 5 in Zinde-Walsh (2008), where for the Cantor distribution it is noted that though $\hat{f}_x(x)$ may diverge, $h^{1-d}\hat{f}_x(x)$ is bounded and bounded away from zero. Then, even though the limit density does not exist by rescaling it is possible to establish the convergence rate of the estimator of the conditional distribution as a functional on $D_0(R^{d_x})$; the rate is $n^{-\frac{1}{2}}h^{1-d}$ and is faster than the root-n rate.

6 Conditional moments

Consider now a conditional moment of a function $g(y)$, of $y \in R^{d_y} : E_{y|x}g(y) = m(x)$, with $m(x)$ measurable with respect to F_x .

When the conditional density function exists in L_1 we write $m(x) = \int g(y)f_{y|x}(x, y)dy$ (assuming that the integral exists). As a generalized function (in x) $m(x)$ can be presented on the space $D(W)$; $W = (0, 1)^{d_x}$ by the value of the functional for ψ :

$$(m, \psi) = \int m(x) \psi(F(x)) dF(x) = \int \left[\int g(y) f_{y|x}(x, y) dy \right] \psi(F(x)) dF(x).$$

To give meaning to (m, ψ) regardless of the existence of the conditional density as a function, $\int g(y)f_{y|x}(x, y)dy$ needs to be characterized as a generalized function on $D(W)$. To make this possible for an arbitrary distribution

on (x, y) that satisfies Assumption 2 the class of functions g is restricted.

Assumption 3. *The function g is continuously differentiable with respect to the differentiation operator ∂^{d_y} .*

Any polynomial function satisfies Assumption 3, and thus conditional mean of y , or conditional variance (if they exist) can be considered. If the function were not to satisfy the differentiability assumption, the class of distributions would need to be correspondingly restricted.

Consider $D(R^{d_y})$ and a locally finite partition of unity on R^{d_y} by a set of suitable functions, "bump" functions from $D(R^{d_y}) : \{\psi_\nu\}$, where $\psi_\nu \in D(R^{d_y})$, $\psi \geq 0$ and $\sum_\nu \psi_\nu(y) \equiv 1$; also any y can belong to support of only a finite number of ψ_ν . See e.g. Gel'fand and Shilov, 1964, v.1, p.142 for a construction.

Then define $(gf_{y|x}, \psi_\nu) = \int g(y)f_{y|x}(x, y)\psi_\nu(y)dy$; under Assumption 3 this expression is (as usual integrating by parts and using boundedness of support of ψ_ν):

$$\int g(y)f_{y|x}(x, y)\psi_\nu(y)dy = (-1)^{d_y} \int F_{y|x}(x, y) \partial^{d_y} (g(y) \psi_\nu(y)) dy. \quad (18)$$

This expression represents a generalized function on $D(W)$ given for any

$\psi \in D(W)$ by

$$\begin{aligned}
& \left(\int g(y) f_{y|x}(x, y) \psi_v(y) dy, \psi \right) \\
&= (-1)^{d_y} \int \int F_{y|x}(x, y) \partial^{d_y} (g(y) \psi_v(y)) dy \psi(F(x)) dF(x) \\
&= (-1)^{d_y + d_x} \int \int F_{x,y}(x, y) \partial^{d_y} (g(y) \psi_v(y)) dy (\partial^{d_x} \psi)(F(x)) dF(x).
\end{aligned}$$

Because the supports of ψ_v and of ψ are bounded and the function being integrated is bounded, the integral exists.

Assumption 4. (Existence of conditional moment). For a partition of unity, $\{\psi_\nu\}$, the sum

$$\Sigma_v \left(\int g(y) f_{y|x}(x, y) \psi_v(y) dy, \psi \right) \quad (19)$$

converges.

Then (19) represents $(m(x), \psi)$ for the generalized function, $m(x) = \Sigma_v \int g(y) f_{y|x}(x, y) \psi_v(y) dy$, on $D(W)$.

Thus

$$m(x) = \int g(y) f_{y|x}(x, y) dy = \Sigma_v (g f_{y|x}, \psi_\nu),$$

where the sum converges.

Then

$$\begin{aligned}\Sigma_v \int g(y) f_{y|x}(x, y) \psi_v(y) dy &= \int g(y) f_{y|x}(x, y) \Sigma_v \psi_v(y) dy \\ &= \int g(y) f_{y|x}(x, y) dy,\end{aligned}$$

in other words interchanging the order of integration and summation is permitted for the terms on the left-hand side of (18) under Assumption 4. However, this is not the case for terms on the right-hand side of (18). For example, if $g(y) = y$, we have $\partial^{d_y}(g(y) \psi_v(y)) = y \psi'_v + \psi_v$, and $\Sigma_v(\partial^{d_y}(g(y) \psi_v(y))) = 1$, but $\int F_{y|x}(x, y) dy$ may not exist.

Thus $(gf_{y|x}, \psi \psi_\nu) =$

$$(-1)^{d_x+d_y} \int \dots \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) \partial^{d_y} [g(y) \psi_v(y_1, \dots, y_{d_y})] dF_x(x) dy_1 \dots dy_{d_y}. \quad (20)$$

Then the conditional moment m as a generalized function on $D(W)$ is given by $(m, \psi) =$

$$\Sigma_v (-1)^{d_x+d_y} \int \dots \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) \partial^{d_y} [g(y) \psi_v(y_1, \dots, y_{d_y})] dF_x(x) dy_1 \dots dy_{d_y} \quad (21)$$

with any $\{\psi_v\}$ representing a partition of unity on R^{d_y} by functions from $D(R^{d_y})$.

7 Limit properties of kernel estimators of conditional mean function.

Suppose that with $d_y = 1$ the conditional mean function $m(x) = E_{y|x}y$ exists; by (21) it then can be represented as

$$\begin{aligned} & (m, \psi) \\ = & \Sigma_v (-1)^{d_x+1} \int \dots \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) [y \psi'_v(y) + \psi_v(y)] dF_x(x) dy_1 \dots dy_{d_y}. \end{aligned} \quad (22)$$

Consider the usual kernel estimator

$$\hat{m}(x) = \frac{\Sigma y_i K\left(\frac{x_i - x}{h}\right)}{\Sigma K\left(\frac{x_j - x}{h}\right)},$$

that can also be represented as

$$\frac{\int y \hat{f}_{x,y}(x, y) dy}{\hat{f}_x(x)} = \frac{\Sigma_v \int y \hat{f}_{x,y}(x, y) \psi_v(y) dy}{\hat{f}_x(x)}.$$

Then for any continuously differentiable $\tilde{\psi}(x)$

$$\begin{aligned} (\hat{m}, \tilde{\psi}) &= \int \frac{\Sigma_v \int y \hat{f}_{x,y}(x, y) \psi_v(y) dy}{\hat{f}_x(x)} \tilde{\psi}(x) dx \\ &= -\Sigma_v \int \frac{\int \partial^{d_x} \hat{F}_{x,y}(x, y) \frac{d}{dy} [y \psi_v(y)] dy}{\hat{f}_x(x)} \tilde{\psi}(x) dx \\ &= -\Sigma_v (\hat{m}, \tilde{\psi} \psi_v). \end{aligned}$$

Consider ψ and $\tilde{\psi} = \psi \hat{f}$; by the Lemma $(\hat{m}, \tilde{\psi} \psi_v) =$

$$\begin{aligned} & (-1)^{d_x+1} \int \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\hat{F}_x(x) \right) \frac{d}{dy} [y \psi_v(y)] d \left(\hat{F}_x(x) \right) dy \quad (23) \\ &= (-1)^{d_x+1} \int \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\hat{F}_x(x) \right) [y \psi'_v(y) + \psi_v(y)] d \left(\hat{F}_x(x) \right) dy. \end{aligned}$$

Assumption 5. The conditional variance $\sigma^2(x) = E_{y|x} y^2$ defines a generalized function on $D(W)$.

Assumption 5 implies that for any $\psi \in D(W)$ the value of the functional $(\sigma^2, \psi) = \int \sigma^2(x) \psi(F_x(x)) dF_x(x)$ is always bounded; this is required to bound the variance for the limit process. By (21) for a partition of unity, $\{\psi_v\}$

$$(\sigma^2, \psi) = \sum_v (-1)^{d_x+1} \int \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) (y^2 \psi_v(y))' dF_x(x) dy.$$

Theorem 4. Suppose that Assumptions 1-5 hold, the bandwidth parameter $h = cn^{-\alpha}$, where $\alpha < \frac{1}{4}$. Then the estimator $\hat{m}(x)$ for a random sample $\{(x_i, y_i)\}_{i=1}^n$ as a generalized random function on $D(W)$ converges at the rate $n^{-\frac{1}{2}}$ to the generalized function m that provides (22); the limit process for $\sqrt{n}(\hat{m} - m)$ on $D(W)$ is given by a $\psi \in D(W)$ indexed random functional

Q_m with $(Q_m, \psi) =$

$$\begin{aligned} & \Sigma_v (-1)^{d_x+1} \int \dots \{ \int U_{x,y} \partial^{d_x} \psi(F_x(x)) dF_x(x) \\ & + \int F_{x,y}(x, y) (\partial^{d_x})^2 \psi(F_x(x)) U_x dF_x(x) \\ & + \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) dU_x \} [y \psi'_v(y) + \psi_v(y)] dy_1 \dots dy_{d_y}, \end{aligned}$$

where $U_x, U_{x,y}$ are Brownian bridge processes with dimension $d_x, d_x + 1$, correspondingly; as a generalized random process the limit process Q_m of $\sqrt{n}(\hat{m} - m)$ is Gaussian with mean functional zero and covariance bilinear functional C , given for any ψ_1, ψ_2 by

$$(C, (\psi_1, \psi_2)) = \text{cov}[(Q_m, \psi_1), (Q_m, \psi_2)].$$

Proof. See Appendix.

Similarly to the kernel estimator for the conditional distribution the conditional mean estimator converges at parametric rate as a functional on $D(W)$ for any distribution. When a positive conditioning density exists it is possible to represent the conditional mean as a functional on $D(R^{d_x})$, by the same arguments as in the Lemma. In the case of Remark 3 a similar rescaling provides a faster convergence rate for the estimator considered as a functional on $D(R^{d_x})$.

8 Conclusion and further questions

The approach employed here makes it possible to avoid any restrictions when defining density, conditional distribution and conditional density as well as conditional moments for a smooth function (e.g. conditional expectation or second moment).

The usual kernel estimators converge to the limit generalized functions at a parametric rate; the limit process is provided by a Gaussian process in the space of generalized functions, that is a Gaussian process indexed by well-behaved functions from the appropriate spaces.

The results here were based on a random sample of observations to simplify exposition; extension to stationary ergodic or mixing processes can be obtained. Further extensions to relax homogeneity and independence are a subject of future research.

The limit results imply that with a judicious selection of indexing functions one could use the kernel estimators for inference in very general situations; this investigation is mostly left for future research.

9 Appendix.

Proof of Theorem 2.

Define a generalized function e_{nhj} such that the value of the functional

for $\psi \in G$ is

$$(e_{nhj}, \psi) = \int \frac{1}{\Pi h_i} K\left(\frac{x - x_j}{h}\right) \psi(x) dx - (f, \psi)$$

and consider $e_{nh} = \frac{1}{n} \sum_{j=1}^n e_{nhj}$; this generalized function provides $\hat{f} - f$.

The expectation functional $E e_{hn}$ gives the generalized bias of the estimator \hat{f} , $Bias(\hat{f})$, see (8).

Next to derive the variance functional consider $T_{lj} = E(e_{nhl}, \psi_1)(e_{hnj}, \psi_2)$.

For $l \neq j$ by independence

$$\begin{aligned} T_{lj} &= E(e_{nhl}, \psi_1)(e_{nhj}, \psi_2) = E(e_{nhl}, \psi_1)E(e_{nhj}, \psi_2) \\ &= \left(Bias(\hat{f}), \psi_1\right) \left(Bias(\hat{f}), \psi_2\right). \end{aligned}$$

For $l = j$

$$\begin{aligned} T_{jj} &= E(e_{nhj}(x), \psi_1)(e_{nhj}(x), \psi_2) \\ &= \int \left[\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_1(x) dx - (f, \psi_1) \right] \times \\ &\quad \left[\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_2(x) dx - (f, \psi_2) \right] dF(x_j) \\ &= T_{jj}^1 + T_{jj}^2, \end{aligned}$$

where

$$T_{jj}^1 = \int \left(\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_1(x) dx \right) \left(\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_2(x) dx \right) dF(x_j)$$

and $T_{jj}^2 =$

$$\begin{aligned} & - \int \left[\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_1(x) dx \right] dF(x_j) \times (f, \psi_1) \\ & - \int \left[\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_2(x) dx \right] dF(x_j) \times (f, \psi_2) \\ & + (f, \psi_1) \times (f, \psi_2). \end{aligned}$$

For every vector h and $s = 1, 2$

$$\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_s(x) dx = \int K(w) \psi_s(x_j - hw) dw.$$

It follows by substituting into T_{jj}^2 and expanding ψ_s that $T_{jj}^2 = -E\psi_1(x) E\psi_2(x) + \bar{h}R_2$.

Similarly,

$$\begin{aligned} T_{jj}^1 &= \int \left(\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_1(x) dx \right) \left(\int \frac{1}{\Pi h_i} K\left(\frac{x_j - x}{h}\right) \psi_2(x) dx \right) dF(x_j) \\ &= \int \left(\int K(w) \psi_1(x_j - hw) dw \int K(w) \psi_2(x_j - hw) dw \right) dF(x_j) = \\ &= \int \left(\int K(w) dw \psi_1(x_j) - \bar{h} \int K(w) \left[\sum_{i=1}^k \frac{\partial \psi_1}{\partial x_i}(x_j - h\tilde{w}) w_i \frac{h_i}{\bar{h}} \right] dw \right) \times \\ &\quad \left(\int K(w) dw \psi_2(x_j) - \bar{h} \int K(w) \left[\sum_{i=1}^k \frac{\partial \psi_2}{\partial x_i}(x_j - h\tilde{w}) w_i \frac{h_i}{\bar{h}} \right] dw \right) dF(x_j) \\ &= E\psi_1(x) \psi_2(x) + \bar{h}R_1; \end{aligned}$$

where after the change of variable $\psi_s(x_j - hw)$ is expanded around the point x_j . Next we establish that $|R_1| < \infty, |R_2| < \infty$.

Indeed,

$$\psi_s(x - hw) = \psi_s(x) - \bar{h} \sum_{i=1}^k \frac{\partial \psi_s}{\partial x_i}(x - h\tilde{w}) w_i \frac{h_i}{\bar{h}}, s = 1, 2, \quad (24)$$

where $\tilde{w} = \alpha w$ for some $0 \leq \alpha \leq 1$ and since $h_i \leq \bar{h}$ and $|w| < 1$ on support of K

$$\left| \sum_{i=1}^k \frac{\partial \psi_s}{\partial x_i}(x - h\tilde{w}) w_j \frac{h_j}{\bar{h}} \right| \leq \left| \sum_{i=1}^k \frac{\partial \psi_s}{\partial x_i}(x - h\tilde{w}) \right|$$

holds and the right-hand side is uniformly bounded by some $B_{\psi_s} < \infty$ since $\psi_s \in D_{l+k}(U)$. Thus

$$|R_1| \leq B_{\psi_1} \sup \psi_2 + B_{\psi_2} \sup \psi_1 + \bar{h} B_{\psi_1} B_{\psi_2}.$$

Similarly, $|R_2| < \infty$.

Combining we get that $T_{jj} = \text{cov}(\psi_1, \psi_2) + O(\bar{h})$ as $\bar{h} \rightarrow 0$.

Consider now

$$\begin{aligned} \eta_{nhj} &= n^{\frac{1}{2}}[e_{nhj} - E(e_{nhj})]; \\ \eta_{nh} &= \frac{1}{n} \sum \eta_{nhj}. \end{aligned} \quad (25)$$

Note that here $\eta_{nhj} = n^{\frac{1}{2}}(e_{nhj} - \text{Bias}(\hat{f}))$. This generalized random function

has expectation zero. In the covariance the terms where $l \neq j$ are zero and

$$\begin{aligned} & n^{-1} E(\eta_{nhj}, \psi_1)(\eta_{nhj}, \psi_2) \\ &= T_{jj} + O(\bar{h}), \end{aligned}$$

and thus converges to $cov(\psi_1, \psi_2)$.

Next (similarly to Zinde-Walsh, 2008) we show that for any set of linearly independent functions $\psi_1, \dots, \psi_m \in D$ with $E(\psi_l^2) > 0$ the joint distribution of the vector

$$\vec{\eta}_{nh} = ((\eta_{nh}, \psi_1), \dots, (\eta_{nh}, \psi_m))'$$

converges to a multivariate Gaussian. Define similarly the vector $\vec{\eta}_{nhj}$ with components (η_{nhj}, ψ_l) . Denote by S the $m \times m$ matrix with ts component $\{S\}_{ts} = (C, (\psi_t, \psi_s))$ where the functional C is given by (9). Denote by \hat{S}_n the covariance matrix of $\vec{\eta}_{nhj}$. By the convergence results for T_{lj} , $\hat{S}_n \rightarrow \Sigma$. Since the functions ψ_1, \dots, ψ_m are linearly independent and $E(\psi_l^2) > 0$ the matrix S and thus \hat{S}_n for large enough n is invertible. Define ξ_{nhj} to equal $\hat{S}_n^{-1/2} \vec{\eta}_{nhj}$, then $\hat{S}_n^{-1/2} \vec{\eta}_{nhj} - S^{-1/2} \vec{\eta}_{nhj} \rightarrow_p 0$.

Next, consider an $m \times 1$ vector λ with $\lambda' \lambda = 1$. The random variables $\lambda' \xi_{nhj}$ are independent with expectation 0, $var \sum \lambda' \xi_{nhj} = 1$; they satisfy the Liapunov condition: $\sum E |\lambda' \xi_{nhj}|^{2+\delta} \rightarrow 0$ for $\delta > 0$ since the kernel function is bounded with finite support. Thus

$$\sum \lambda' \xi_{nhj} \rightarrow_d N(0, 1)$$

and by the Cramer-Wold theorem convergence to a limit Gaussian process for $\hat{S}_n^{-1/2} \overrightarrow{\eta}_{nh}$ and thus for $S^{-1/2} \overrightarrow{\eta}_{hn}$ follows. ■

Proof of Theorem 3.

Since for a smooth kernel $\hat{F}(x, y) \in \Phi_c$ by the Lemma the value of the functional for $\psi \in D(0, 1)^{d_x}$, $(\hat{F}_{y|x}, \psi)$ is the same as $(\hat{F}_{y|x}, \tilde{\psi})$, with the latter defined by (13) where $\tilde{\psi} = \hat{f}_x \psi(\hat{F}_x)$. Thus for any $\psi \in D(0, 1)$:

$$\begin{aligned} & (\hat{F}_{y|x}, \psi) \\ = & (-1)^{d_x} \int \frac{1}{n} \Sigma \bar{G} \left(\frac{y - y_i}{h_y} \right) \bar{K} \left(\frac{x_i - x}{h} \right) \partial^{d_x} \psi \left(\Sigma \bar{K} \left(\frac{x_i - x}{h} \right) \right) d \left(\Sigma \bar{K} \left(\frac{x_i - x}{h} \right) \right) \end{aligned} \quad (26)$$

More concisely it is $(\hat{F}_{y|x}, \psi) =$

$$\begin{aligned} & (-1)^{d_x} \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\hat{F}_x(x) \right) d \left(\hat{F}_x(x) \right) \\ & + (-1)^{d_x} \left[\int \tilde{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\tilde{F}_x(x) \right) d \left(\tilde{F}_x(x) \right) - \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\hat{F}_x(x) \right) d \hat{F}_x(x) \right]. \end{aligned}$$

Here "hat" indicates empirical distribution function and "tilde" the kernel estimated distribution function. By standard arguments the smooth kernel introduces a bias; by the usual expansions using differentiability of ψ we get that for the second order kernel

$$\begin{aligned} & (-1)^{d_x} \left[\int \tilde{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\tilde{F}_x(x) \right) d \left(\tilde{F}_x(x) \right) - \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi \left(\hat{F}_x(x) \right) d \hat{F}_x(x) \right] \\ = & O_p(h^2). \end{aligned}$$

Represent $(-1)^{d_x} \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi(\hat{F}_x(x)) d(\hat{F}_x(x))$ as

$$\begin{aligned}
& (-1)^{d_x} \left\{ \int F_{x,y} \partial^{d_x} \psi(F_x) d(F_x) + \int F_{x,y} [(\partial^{d_x} \partial^{d_x} \psi)(F_x) (\hat{F}_x - F_x) + r (\hat{F}_x - F_x)^2] d(F_x) \right. \\
& + \int F_{x,y} \partial^{d_x} \psi(F_x) d(\hat{F}_x - F_x) + \int F_{x,y} (\partial^{d_x} \partial^{d_x} \psi)(\tilde{F}_x) (\hat{F}_x - F_x) d(\hat{F}_x - F_x) \\
& + \int (\hat{F}_{x,y} - F_{x,y}) \partial^{d_x} \psi(F_x) dF_x + \int (\hat{F}_{x,y} - F_{x,y}) \partial^{d_x} \psi(F_x) d(\hat{F}_x - F_x) \\
& + \int (\hat{F}_{x,y} - F_{x,y}) (\partial^{d_x} \partial^{d_x} \psi)(\tilde{F}_x) (\hat{F}_x - F_x) dF_x \\
& \left. + \int (\hat{F}_{x,y} - F_{x,y}) (\partial^{d_x} \partial^{d_x} \psi)(\tilde{F}_x) (\hat{F}_x - F_x) d(\hat{F}_x - F_x) \right\}
\end{aligned}$$

where \tilde{F}_x represents an intermediate value and takes values in $(0, 1)^{d_x}$; by properties of $\psi \in D(W)$ the function $(\partial^{d_x} \partial^{d_x} \psi)(\tilde{F}_x)$ is bounded. Then $\sqrt{n}(\hat{F}_{y|x} - F_{y|x}, \psi)$ can be expressed as

$$Q_\psi \left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{xy} - F_{xy}) \right) + n^{-\frac{1}{2}} R \left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{xy} - F_{xy}) \right),$$

where

$$\begin{aligned}
& Q_\psi \left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{xy} - F_{xy}) \right) \\
& = \int F_{x,y} [(\partial^{d_x} \psi)(F_x)] d\sqrt{n}(\hat{F}_x - F_x) + \int \sqrt{n}(\hat{F}_{x,y} - F_{x,y}) [(\partial^{d_x} \psi)(F_x)] dF_x \\
& + \int F_{x,y} [(\partial^{d_x} \partial^{d_x} \psi)(F_x)] \sqrt{n}(\hat{F}_x - F_x) d(F_x)
\end{aligned}$$

and $R(., .)$ is a bounded function.

Since the limit process of $\sqrt{n}(\hat{F}_x - F_x)$ is $U_.$, a Brownian bridge, and the

function Q_ψ is continuous in its arguments, by Donsker's theorem we can express the limit process for $\sqrt{n} \left(\hat{F}_{y|x} - F_{y|x}, \psi \right)$ as $(Q_{y|x}, \psi) = Q_\psi(U_x, U_{xy})$ by substituting the limit Browning bridge processes for the arguments of $Q_\psi(\cdot, \cdot)$.

For any $\psi_1, \dots, \psi_l \in D(W)$ the joint limit process for

$$\sqrt{n} \left(\hat{F}_{y|x} - F_{y|x}, \psi_1 \right), \dots, \sqrt{n} \left(\hat{F}_{y|x} - F_{y|x}, \psi_l \right)$$

is similarly given by the joint process of $Q_{\psi_1}(U_x, U_{xy}), \dots, Q_{\psi_l}(U_x, U_{xy})$. This is a Gaussian process. The mean is zero since Q_ψ is linear in its arguments and the covariance is given by $\text{cov}(Q_{\psi_1}(U_x, U_{xy}), Q_{\psi_2}(U_x, U_{xy})) = \text{cov}((Q_{y|x}, \psi_1), (Q_{y|x}, \psi_2))$. Existence follows from boundedness of the functions in the expressions and bounded support of ψ .

By assumption of the theorem $h^2 = o(n^{-\frac{1}{2}})$, thus the limit process is fully described by $Q_{y|x}$.

■

Proof of Theorem 4.

For (23) we obtain

$$\begin{aligned}
& (-1)^{d_x+1} \int \int \hat{F}_{x,y}(x, y) \partial^{d_x} \psi(\hat{F}_x(x)) [y\psi'_v(y) + \psi_v(y)] d(\hat{F}_x(x)) dy \\
= & (-1)^{d_x+1} \{ \int \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) [y\psi'_v(y) + \psi_v(y)] d(F_x(x)) dy \\
& + \int \int [\hat{F}_{x,y}(x, y) - F_{x,y}(x, y)] [y\psi'_v(y) + \psi_v(y)] \partial^{d_x} \psi(F_x(x)) d(F_x(x)) dy \\
& + \int \int F_{x,y}(x, y) (\partial^{d_x})^2 \psi(F_x(x)) [\hat{F}_x(x) - F_x(x)] [y\psi'_v(y) + \psi_v(y)] d(F_x(x)) dy \\
& + \int \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) [y\psi'_v(y) + \psi_v(y)] d(\hat{F}_x(x) - F_x(x)) dy \\
& + \tilde{R} \},
\end{aligned}$$

where \tilde{R} combines the remaining terms. Analogously to the proof of Theorem

3 $\sqrt{n}(\hat{m} - m, \psi\psi_v)$ is represented as

$$Q_{\psi\psi_v} \left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{xy} - F_{xy}) \right) + n^{-\frac{1}{2}} R \left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{xy} - F_{xy}) \right).$$

The limit process for the first functional is expressed via a value of the functional for Brownian bridges,

$$\begin{aligned}
Q_{\psi\psi_v}(U_x, U_{xy}) = & \int \int U_{x,y} [y\psi'_v(y) + \psi_v(y)] \partial^{d_x} \psi(F_x(x)) d(F_x(x)) dy \\
& + \int \int F_{x,y}(x, y) (\partial^{d_x})^2 \psi(F_x(x)) U_x [y\psi'_v(y) + \psi_v(y)] d(F_x(x)) dy \\
& + \int \int F_{x,y}(x, y) \partial^{d_x} \psi(F_x(x)) [y\psi'_v(y) + \psi_v(y)] d(U_x) dy.
\end{aligned} \tag{28}$$

This process is Gaussian with mean zero; summing over v we get a

zero mean limit process, $(Q_m, \psi) = \sum_v Q_{\psi\psi_v}(U_x, U_{xy})$. We need to verify that the bilinear covariance functional $cov((Q_m, \psi_1), (Q_m, \psi_2))$ is well-defined (bounded) for any ψ_1, ψ_2 .

Since expectation of Q_m is zero

$$\begin{aligned} |cov((Q_m, \psi_1), (Q_m, \psi_2))| &\leq [E(Q_m, \psi_1)^2 E(Q_m, \psi_2)^2]^{\frac{1}{2}}, \\ E(Q_m, \psi)^2 &= E\left(\sum_v Q_{\psi\psi_v}(U_x, U_{xy})\right)^2. \end{aligned}$$

Thus it is sufficient to consider variances for some ψ .

The representation in (28) involves three terms, it is sufficient to show that the variance of the sum of each type of term over all v is bounded.

Recall that here $cov(U_{z_1}, U_{z_2}) = F(\tilde{z}) - F(z_1)F(z_2)$, where $\tilde{z} = z_1 \wedge z_2$.

Start with the first term in (28) and consider its variance.

Evaluate

$$\begin{aligned}
& E\left\{\int \dots \int U_{x_1, y_1} U_{x_2, y_2} [y_1 \psi'_{v_1}(y_1) + \psi_{v_1}(y_1)] [y_2 \psi'_{v_2}(y_2) + \psi_{v_2}(y_2)] dy_1 dy_2 \right. \\
& \quad \cdot \partial^{d_x} \psi(F_x(x_1)) d(F_x(x_1)) \partial^{d_x} \psi(F_x(x_2)) d(F_x(x_2)) \left. \right\} \\
& = E_1 - E_{1,2} \text{ with} \\
E_1 & = \left\{ \int \dots \int F(x_1, y_1) [y_1 \psi'_{v_1}(y_1) + \psi_{v_1}(y_1)] \left[\int^{y_1} [y_2 \psi'_{v_2}(y_2) + \psi_{v_2}(y_2)] dy_2 \right] dy_1 \right. \\
& \quad \cdot \partial^{d_x} \psi(F_x(x_1)) d(F_x(x_1)) \int^{x_1} \partial^{d_x} \psi(F_x(x_2)) d(F_x(x_2)) \left. \right\} \\
\text{and } E_{1,2} & = \tilde{E}_1 \tilde{E}_2 \text{ where for } i = 1, 2 \\
\tilde{E}_i & = \int \dots \int F(x_i, y_i) [y_i \psi'_v(y_i) + \psi_v(y_i)] \partial^{d_x} \psi(F_x(x_i)) d(F_x(x_i)) dy_i.
\end{aligned}$$

For E_1 integrating we get (dropping the subscript 1 on variables)

$$\int \dots \int F(x, y) [y^2 \psi'_{v_1}(y) \psi_{v_2}(y) + y \psi_{v_1}(y) \psi_{v_2}(y)] dy \cdot \frac{1}{2} \partial^{d_x} \psi^2(F_x(x)) dF(x).$$

By construction of the partition of unity $|\Sigma \psi'_{v_1}(y) \psi_{v_2}(y)|$ as well as $\Sigma \psi_{v_1}(y) \psi_{v_2}(y)$ are uniformly bounded, say both by some \tilde{B} . We get

$$\begin{aligned}
& \left| \int \dots \int F(x, y) [y^2 \psi'_{v_1}(y) \psi_{v_2}(y) + y \psi_{v_1}(y) \psi_{v_2}(y)] dy \cdot \frac{1}{2} \partial^{d_x} \psi^2(F_x(x)) dF(x) \right| \\
& \leq \frac{\tilde{B}}{2} [(E_{|x}(y^2), \psi^2) + |(E_{|x} y, \psi^2)|].
\end{aligned}$$

Note that $\psi^2 \in D(W)$. By Assumption 5 then this contribution to the covariance is bounded.

Similarly boundedness of the other contributions from all the terms into the covariance can be obtained. By the condition $h^2 = o(n^{-\frac{1}{2}})$ on the bandwidth the bias does not affect the limit process.

■

References

- [1] Anderson, G., O. Linton and Y.-J. Whang (2012) Nonparametric estimation and inference about the overlap of two distributions, *Journal of Econometrics*, 171, pp. 1-23.
- [2] Azzalini, A. (1981). A note on the estimation of the distribution function and quantiles by a kernel method. *Biometrika*, 68 326-328.
- [3] Carrasco, M., J.-P. Florens, and E. Renault (2007) Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization, *Handbook of Econometrics*, Vol. 6B, ed. by J.Heckman and E. Leamer. Amsterdam: North-Holland.
- [4] Carrasco, M. and J.-P. Florens (2010), A Spectral Method for Deconvolving a Density, *Econometric Theory*, 27 , pp. 546-581.
- [5] Chang, J.C. and D. Pollard (1997) Conditioning as Disintegration, *Statistica Neerlandica*, 51, pp.287-317.

- [6] Devroye, L. and L. Györfi (1985) Nonparametric Density Estimation: The L_1 View, Wiley, New York.
- [7] Gel'fand, I.M. and G.E.Shilov (1964) Generalized Functions, Vol.1, Properties and Operations, Academic Press, San Diego.
- [8] Gel'fand, I.M. and G.E.Shilov (1964) Generalized Functions, Vol.2, Spaces of Test functions and Generalized Functions, Academic Press, San Diego.
- [9] Gel'fand, I.M. and N.Ya Vilenkin (1964) Generalized Functions, Vol.4, Applications of Harmonic Analysis, Academic Press, San Diego.
- [10] Komlos, J., Major, P. and Tusnady, G. (1975) An approximation of partial sums of independent rv's and the sample df. I, Wahrsch verw Gebiete/Probability Theory and Related Fields, 32, 111–131.
- [11] Komlos, J., Major, P. and Tusnady, G. (1976) An approximation of partial sums of independent rv's and the sample df. II, Wahrsch verw Gebiete/Probability Theory and Related Fields, 34, 33–58.
- [12] Li, Q. and J.Racine (2007) Nonparametric Econometrics: theory and practice, Princeton University Press.
- [13] Lu, Z.-Q. (1999) Nonparametric regression with singular design, Journal of Multivariate analysis, 70, 177-201.

- [14] Pagan, A. and A. Ullah (1999) Nonparametric Econometrics, Cambridge University Press.
- [15] Phillips, P.C.B. (1991) A shortcut to LAD estimator asymptotics, *Econometric Theory*, 7, 450-463.
- [16] Pfanzagl, P. (1979) Conditional Distributions as Derivatives, *The Annals of Probability* , Vol. 7, pp. 1046-1050.
- [17] Schwartz, L. (1966) "Théorie des distributions", Hermann, Paris.
- [18] Sklar, A. (1973), Random variables, joint distributions, and copulas. *Kybernetika* 9, 449-460.
- [19] Sobolev, S.L. (1992) Cubature Formulas and Modern Analysis. Gordon and Breach Science Publishers.
- [20] Zinde-Walsh, V. (2008) Kernel Estimation when Density May not Exist, *Econometric Theory*, 24, pp. 696-725.
- [21] Zinde-Walsh, V. (2011), Presidential Address: Mathematics in economics and econometrics, *Canadian Journal of Economics*, v.44, pp. 1052-1068.